DOCUMENT RESUME

ED 327 586 TM 016 043

AUTHOR Murchan, Damian P.
TITLE Essay versus Objective Achievement Testing in the Context of Large-Scale Assessment Programs.
PUB DATE Mar 89
NOTE 22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 28-30, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Comparative Testing; Construct Validity; Content Validity; Curriculum Evaluation; Error of Measurement; *Essay Tests; Foreign Countries; Geography; Graduation Requirements; Multiple Choice Tests; *Objective Tests; Secondary Education; *Secondary School Students; *Test Format; Testing Programs; Test Reliability
IDENTIFIERS *Intermediate Certificate Examination; Ireland; *Large Scale Programs

ABSTRACT

The reliability, content validity, and construct validity were compared for two test formats in a public examination used to assess a secondary school geography course. The 11-item geography portion of the Intermediate Certificate Examination (essay examination) was administered in June 1987 to 400 secondary school students in Ireland who also took an objective 43-item multiple-choice geography test about 7 months later (in November 1987). The scores on the essay test were less reliable and were susceptible to a significantly larger standard error of measurement than were those on the objective test. Test format was also found to have an effect on classroom practices. The two tests measured different skills, and there was insufficient evidence to prove that writing ability or quality of handwriting influenced scores assigned to essays. It is suggested that a mixed test format is a desirable way to measure achievement in content areas. Five tables present study data. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED327586

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

DAMIAN P. MURCHAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Essay Versus Objective Achievement Testing in the Context of Large-Scale Assessment Programs

Damian P. Murchan
Cornell University

Paper Presented at the 1989 Annual Conference of the National Council on Measurement in Education, San Francisco, California.

TM016043

Essay Versus Objective Achievement Testing in the Context of Large-Scale Assessment Programs

Damian P. Murchan

Cornell University

This paper compared the reliability, content validity, and construct validity of two test formats in a public examination used to assess a secondary school geography course. Results indicated that scores on the essay exam were less reliable and susceptible to a significantly larger standard error of measurement than an objective test of equal length. The test format used was found to have a tremendous influence on classroom practices. The two tests measured different skills, both of them assessing higher and lower-order abilities, and there was insufficient evidence to prove that writing ability or quality of handwriting influenced scores assigned to essays. The lessons for designers of large-scale assessment programs are that a mixed test format containing essays and objective questions is a desirable way to measure achievement in content areas. Over-reliance on either item type will lessen the quality and quantity of inferences that can be made from scores obtained.

Background to Study

Over the past few years, State Education Departments nationwide have been examining the role of direct measures of writing in their assessment programs. Lately

---

The author would like to thank Jason Millman for providing invaluable assistance during the course of this study and for his useful comments on an earlier version of this paper.

the emphasis has been switching somewhat to examining how essays can be used to determine student achievement in a wide variety of subject areas. Conversely, most countries of the European Community, many Asian nations and countries of the British Commonwealth are presently testing the lukewarm waters of objective testing, once the anathema of the Old World education systems, lured by the much-heralded superior reliability of such tests. Despite the interest being shown in the comparability of essay and objective methods of assessing pupil achievement, solid evidence is still thin on the ground. We have been inundated with a plethora of studies dealing with inter-rater reliability in grading essays, the findings of which are as diverse as the countries in which they took place. A common conception, among European educators anyway, seems to be that though rater reliability is undeniably not as high as that obtained by objective tests of achievement, the benefit to be gained in terms of measuring higher cognitive skills makes the practice worthwhile. An issue that has escaped close attention however, is that of the test reliability of both essay and objective instruments. In rectifying this deficiency, my study suggests that the issue might not be as dead as Hogan (1981) would have us believe.

Objectives and Methodology

Two primary objectives underpin this study: (i) how reliably can essay and objective achievement tests of the same length measure the same content, and (ii) what exactly is measured by each test? The latter included ascertaining whether essay and objective tests can measure higher level skills, and if so, what these skills are and how relevant are they to the domain of interest. This analysis of the construct validity of the test types also probed the effect of writing ability on the test scores.

Many studies of essay examinations have pronounced sentence on the basis of low inter-rater reliability or poor equating of test forms, their objective counterparts being aquitted of these crimes (Chase, 1986; Marshall, 1967; Marshall and Powers, 1969; Tollefson and Tracy, 1980). This study, in keeping with current practice, has gone beyond questioning these drawbacks and has focused on the core of the concerns, the test reliability and relative construct validity of essay and objective examinations used in a nationwide testing program.

The Intermediate Certificate Examination (ICE), developed by the national Department of Education in the Republic of Ireland, is used to evaluate pupil performance after the first three years of secondary education in approximately 30 subject areas such as English, Mathematics, Science, Economics, French, History, and Geography to name a few. Students typically present 8 or 9 subjects for examination. The geography component of this battey is the topic of interest for this study.

The Intermediate Certificate Geography Exam is completely revamped for each administration. It is a closed-book examination consisting of 11 questions of which examinees answer 5 or 6, depending on the particular course followed in school. Of the 11 questions, 10 are fundamentally of the essay variety. Students pursuing Syllabus I (sections A, B, and C) have 2 hours in which to take the exam. Syllabus II students answer one additional question from section D and receive 30 extra minutes testing time. The layout of the 1987 geography exam is presented in Table 1

Table 1

Structure of the 1987 ICE Geography Paper

With Corresponding Objective Questions Set for This Investigation

SECTION	QUESTION NUMBER	CONTENT OF QUESTION	MARKS PER QUESTION	NUMBER OF OBJECTIVE ITEMS
A	1	General	36	0
B	2	Ordnance Survey Map	36	4
	3	Town Plan - Mullingar	36	4
	4	Glaciation	36	4
	5	Savanna climate	36	4
C	6	Ireland:General & West	36	4
	7	Ireland: Fishing	36	0
	8	EC: General	36	4
	9	EC: Italy	36	0
D	10	North America:	50	0
	11	South America	50	0

Notes: 1. All students answered Sections A, B, and C. Students were allowed to choose to answer two of four essays in each of Sections B and C. Those students pursuing Syllabus II also answered one question from Section D.

2. EC = European Community.

Examination papers are assigned a letter grade using the cut scores shown in Table 2 (Department of Education, 1984). Grades of A, B or C denote Honors, a grade of D constitutes a Pass, while anything below a D is considered Failure in that subject. The Irish Department of Education, just as do their European counterparts, go to tremendous lengths to minimize marker bias or error in the grading of scripts written by ICE candidates. The names of the people who set ICE papers are not published (Madaus & Macnamara, 1970). They are generally set by the inspectorate of the

Table 2
Letter grades and percentage range
equivalents on the Intermediate Certificate Examination

LETTER GRADE	PERCENTAGE RANGE
A	85 or over
B	70 but less than 85
C	55 but less than 70
D	40 but less than 55
E	25 but less than 40
F	10 but less than 25
No Grade	Less than 10

Department and administered over a period of approximately 3 weeks beginning on the second Wednesday in June each year (Department of Education, l984). Students do not write their names anywhere on the paper, instead identifying themselves solely by means of a 4 or 5 digit exam number supplied to them by the Department. Scorers are usually teachers of the subject being examined, no teacher receiving scripts written by students in his/her school. Marking keys are prepared for each paper and prior to beginning the grading process, all examiners come together in conference and any problems with the marking scheme are discussed, and, if necessary, the scheme is modified. Examiners then proceed to grade the papers, under the supervision of Departmental inspectors who review a small sample of scripts from each examiner in an attempt to ensure proper standards. The distribution of each examiner's grades is checked against the distribution for examiners in general. Deviations from this general distribution receive further scrutiny (Greaney & Kellaghan, l984).

A sample of 400 students was selected to participate in this study. These students, whose average age was 15 years and 7 months, were located in 6

secondary schools in Ireland. In June 1987, the students took the ICE. In November 1987, an objective test was administered to students in the sample. The objective instrument was a 43-item, multiple-choice test. Of these items, 24 covered 6 of the specific topics included on the ICE taken by students in June 1987. 16 items covered 8 topics listed on the official geography syllabus which were not subsumed by essay items appearing on the 1987 ICE. These 40 geography items were designed to assess both higher and lower level thinking skills. The final 3 items dealt with student attitude to the two test formats. Further data were gathered from students using a short essay on a neutral English topic. This essay yielded two scores: an assessment of the physical layout and attractiveness of the writing sample (handwriting, neatness and general appearance) in addition to a measure of the level of writing ability displayed. The latter category included writing style, word choice and fluidity of writing. Included also was a 46-item questionnaire that asked teachers for information on what topics they had stressed in their teaching as well as their professional consideration of issues related to testing and to this study in particular. Furthermore, informal meetings were conducted with teachers and students on a variety of issues connected with this project. Readers interested in examining instruments used in this study are referred to Murchan, 1989.

Conclusions

Reliability of the Tests

The geography component of the ICE posed many interesting problems in relation to ascertaining its reliability. The primary hurdle was the wide choice given to students in choosing questions. Numbers of examinees opting for different

questions ranged from 322 students for essay 2 (Ordnance Survey) to only 19 for essay 5 (Savanna Climate). Conventional statistical packages offer little by way of a defensible reliability algorithm in such a situation.

The primary method chosen to obtain an overall reliability figure for the 8 essays in Sections B and C involved using weighted Z scores according to the formula proposed by Fisher, conceptualizing the exam as a composite of two scores. Crocker and Algina (1986, page 505) provide a rationale for utilising the formula shown below.

$$\text{Rel 1} = \frac{r_{BB}\, s_B^2 + r_{CC}\, s_C^2 + 2r_{BC}\, s_B\, s_C}{s_B^2 + s_C^2 + 2r_{BC}\, s_B\, s_C} \qquad (1)$$

where

Rel 1 is the estimated reliability of the 4 essays answered in Sections B &C,

r_{BB} and r_{CC} are estimates of the intra-section reliabilities

S^2_B is the variance of the total scores for Section B,

S^2_C is the variance of the total scores for Section C,

r_{BC} is the correlation between the total scores for Section B and Section C,

S_B is the standard deviation of the total scores for Section B, and

S_C is the standard deviation of the total scores for Section C

This method yielded a reliability coefficient of .79 for Sections B and C. Two additional analyses using adjusted correlations resulted in coefficients of .71 and .79, though the procedures employed were not as sophisticated as for the primary method.

Calculation of reliability figures for the objective test as a whole used both KR 20 and Split-Half procedures, the latter being the better estimate because of the differing cognitive nature of the questions (measuring both higher- and lower-order skills). Reliability data for different subsections of the objective test are presented in Table 3.

Table 3

KR20 and Split-Half Reliabilities of Objective Test

Scale	KR20	Split-Half
All 40 items	.71	.77
24 items based on ICE essays	.54	.59
16 items not based on ICE essays	.60	.65
20 lower-order items	.53	.53
20 higher-order items	.56	.56
12 lower-order ICE items	.39	.41
12 higher-order ICE items	.29	.28
8 lower-order non-ICE items	.26	.25
8 higher-order non-ICE items	.51	.51

n = 362

The obtained coefficient of .77, when adjusted to match the length of sections B and C of the ICE became .89, which is greater than the .79 figure for the ICE. Pilot-testing of the items would undoubtably have increased this gap even further. An additional difference between the coefficients for the two tests is that we can have confidence in the objective test reliability coefficient whereas we know that the ICE figure is an overestimate since inter-rater unreliability is not considered.

Standard errors of measurement for both tests are markedly different also, the greater error being evident in the case of ICE scores. According to data presented in this study, students presenting Syllabus I geography for examination in the ICE and scoring in the exact middle of a grade category have, due to a standard error of 7 percentage points, a 28% chance of receiving a grade differing by 1 grade level up or down if they were to take a parallel exam a second time. In contrast, the probability of

test unreliability affecting examinees' grades in a retest if an objective test format were employed is less than 4%. Furthermore, the former figure, when coupled with Madaus's (1970) calculation of a standard error of 2.9%, due to inter-rater unreliability, for grades assigned on Leaving Certificate geography suggests a total standard error of 7.6% for scores assigned to students on the ICE. This leads to the ominous conclusion, using normal probability tables, that for a student obtaining a score in the middle of a grade band, there is a 33% chance that the examinee's score would deviate up or down by one grade level if the student were retested using a parallel Intermediate Certificate geography exam graded by a different scorer. In other words, if a student gets 47.5% on the test, thereby obtaining a grade of D, there is a 1 in 3 chance that the grade assigned on a retest could be a C or an E, the former bestowing honours distinction, the latter being a failing grade. This is a significantly greater probability than the 3.6% chance of such fluctuation occurring with the objective test.

Though Madaus's figure applied to the 1967 Leaving Certificate geography exam, it can be assumed that, given the extra importance attached to grades assigned to students in this secondary school graduation exam, scorers are at least as careful, if not more so, in grading scripts. Therefore, the standard error of 2.9% is probably a conservative estimate to use for the Intermediate Certificate Examination where the stakes involved in mis-classifying a student regarding grade are less than for the LCE.

The above figures have all assumed that an examinee's observed score on the exam falls roughly in the middle of a grade band, say, 62.5% which is a C. However, if a student gets a score close to the cut point between two letter grades on the ICE (40%, 55%, 70%, etc.), combined test and rater unreliability means that there is almost a 1 in 20 chance of obtaining a grade 2 letters away from the original grade on a retest using similar items. So, for the student who just manages to get a C (55%), there is a

5% chance of obtaining a B or an E on a retest! The chance of this happening with the objective test is negligible. Though these figures rely somewhat on an inter-rater reliability study conducted almost 2 decades ago, they do highlight the point that, despite the best efforts to guard against them, serious problem do exist with scores assigned using the essay format .

Validity of the Tests: Content Validity

Little has been said or written about the content validity of essay-based public examinations such as the ICE. The general consensus seems to be that it is preferable to measure a relatively small proportion of the syllabus in detail rather than attempting to do a sweeping assessment of the entire course, as would be the case with an objective test. Findings in this study indicate that, for an essay test, the ICE does remarkably well in relation to domain coverage where students in 1987 were asked questions relating to over half of the syllabus. Much of this coverage may be attributed to Section A, a quasi-objective test consisting of 15 completion-type items.

The examination format employed does have a direct bearing on classroom activity. Data gathered from teachers confirmed the common conception that teachers base a sizeable proportion of their teaching on certain predictions about what is likely to be asked on the ICE. Cases in point are the extensive treatment given to Italy prior to the 1987 exam - most likely due to the fact that it had not been a major question since 1983. The unwritten rule between the Department of Education and teachers governing Section C that there will always be two questions on Ireland and two on the European Community (EC) seemingly prompts many teachers (40% according to my data) to concentrate exclusively on either Ireland or the EC, hoping that their students will be able to answer whatever two items are asked on the broad topic. So, though

the ICE itself may indeed contain questions on an acceptable proportion of the syllabus, the format of the exam along with the teaching strategies employed as a result, mean that students learn only a certain proportion of the syllabus to begin with, an assertion that is borne out by the fact that students do actually choose questions on the topics that their teachers have prepared them for. This whole issue warrants further attention. For example, would reducing the choice of which essays to respond to help alleviate the problem of not teaching the entire syllabus?

In most education systems, especially where large-scale public exams are used to assess student achievement, it is to be expected and indeed it is quite proper and necessary that the test influences the curriculum. Individual responses to the questionnaire in this study show very clearly that most teachers plan their own syllabi based on some unwritten but very well tested and tried rules governing the structure of the Intermediate Certificate Examination paper. In effect the test has redefined the syllabus which no longer matches the domain of the official Departmental syllabus.

Validity of the Tests: Construct Validity

In comparing the constructs being measured by both test formats, topic-to-topic correlations were computed between pairs from 5 topics (Ordnance Survey, Town Plan, Glaciation, Ireland: General, and EC: General). The topic "Savanna Climate" was not included in most analyses due to the fact that only 19 students attempted it on the ICE. As it proved impossible to ascertain in any meaningful way the reliability for each ICE essay due to the choice given to examinees, essay combinations were paired, resulting in correlations being computed between the sum of 2 essays and their 8 corresponding objective items. This pairing of essay topics facilitated the

calculation of alpha reliabilities for the essay composites. Uncorrected and corrected correlations for these combinations are presented in Table 4.

Table 4
Correlations Between
ICE Essay Combinations and Objective Totals

ICE Essays	N[1]	Reliability		Correlation with Objective ITems	
		Essay[2]	Objective[3]	Uncorrected	Corrected
2, 3	176	.61	.30	.19	.44
2, 4	140	.49	.32	.20	.49
3, 4	28	.73	.30	.22	.44
6, 8	33	.81	.17	.32	.76

[1] = Number of students answering both essays
[2] = Alpha reliability of 2 essays
[3] = Spearman-Brown split-half reliability of 8 objective items

The weighted average correlation between the ICE and the objective test, corrected for unreliability of the measures, was found to be .48, thus implying that they are measuring somewhat different attributes of students' skills in geography.

Discovering the nature of the difference in skills measured by the two test formats then switched to testing the hypothesis that it was due to one of the tests measuring predominately higher-order skills whereas the other test measured lower-order skills. Summing students' scores on the four essays answered by them yielded a score for each student on the ICE and this was then correlated with subsections of the objective test in order to get a rough measure of how well the skills required to answer each test correlate. Corrected correlations are shown in Table 5.

Accepting the idea that essays adequately assess higher-level cognitive skills, then the correlation of .72 between the ICE and corresponding higher-order items supports the claim that essays are measuring more strongly higher-order than lower-order skills. However, no significant difference was found between uncorrected correlations between the ICE and objective lower and higher-order items.

Table 5
Corrected Correlation of ICE with Objective Test Subsections

Objective Test Subsections	ICE
24 objective items based on ICE	.55
16 objective items not based on ICE	.60
12 lower-order items based on ICE	.53
12 higher-order items based on ICE	.72

n = 362

It is interesting to note also from Table 5 that the skills underlying the 16 non-ICE objective items correlate somewhat higher with the ICE than do the 24 items designed to match the essay test, though the significance of the difference has not been calculated. If the instrument had been pilot tested on a sample of the population prior to its administration, this anomaly might have been averted.

Having failed to conclusively prove that the difference between the two tests could be explained as their measuring different cognitive abilities, the emphasis shifted to examining the differential effects of writing ability on the tests. Simple correlations computed between each ICE essay and writing scores proved inadequate due to the unreliability of items used in their calculation, though they do at least

indicate a general trend which can be better explained using corrected correlations between larger units of both tests as was done in comparing the two formats. The rough pattern that emerged from this latter analysis was that the ICE essays correlate more strongly with both writing dimensions than do the corresponding objective items, though even in the case of ICE essays and writing, the correlations were quite low. Uncorrected and corrected correlations between both test formats (using combinations of 2 essays and 8 objective items) and scores on both writing dimensions are presented in Table 6.

These data clearly indicate a tendancy for essay scores to correlate more highly with writing scores than do objective scores. The ICE correlates .5 with the writing style dimension and .35 with writing appearance. Corresponding figures for the objective test are .22 and .08 respectively. In other words, using the coefficient of determination we find that 25% of the variance in students' ICE scores is associated with variance on their writing scores. Common variance between scores on the objective test and style scores is negligible. However, the effect of examinees' handwriting on ICE scores, though significant, is not as great as had been thought, the greater effect being recorded for complexity of language, syntax structure and vocabulary level used by the student. Nevertheless, factor analysis of the ICE and writing scores yielded two reasonably distinct factors, with geography essays loading heavily on the first factor thus indicating that writing ability is not a major contaminating factor in scores assigned to students on the ICE.

Table 6
Correlation of Test Formats
With Measures of Writing Ability

Topics	N	Test Format	Writing Appearance U[1]	Writing Appearance C[2]	Writing Style U[1]	Writing Style C[2]
2 & 3	176	Essay[3]	.30	.38	.43	.55
		Objective[4]	.02	.04	.10	.18
2 & 4	140	Essay	.18	.36	.34	.48
		Objective	.07	.12	.11	.19
3 & 4	28	Essay	.25	.29	.35	.41
		Objective	-.12	-.22	.01	.02
6 & 8	33	Essay	.19	.21	.30	.33
		Objective	.14	.34	.26	.63
Weighted Average of Essay[5]			.24	.35	.38	.50
Weighted Average of Objective[5]			.04	.08	.11	.22

[1] = Uncorrected Correlation
[2] = Corrected Correlation
[3] = Sum of students' scores on 2 essays
[4] = Sum of students' scores on 8 objective items
[5] = Averages computed using Fisher's Z transformation of r. (in r units)

Part correlations were also computed for each of the 4 pairs as outlined already in Table 4. Results are presented in Table 7. There is little significant change in correlations between essays and objective items when the effect of writing is partialled out. The weighted average correlations adjusted for the effect of writing appearance and writing style are .4 and .42 respectively. These results support those of the factor

Table 7

Part Correlations Between ICE and Objective Items with Effect of Writing Partialled Out of the Essay Scores

ICE Essays	Objective Items		
	Appearance[1]	Style[2]	Unpartialled[3]
2 & 3	.30	.41	.44
2 & 4	.48	.45	.49
3 & 4	.53	.47	.44
6 & 8	.70	.58	.76

[1] = r between essays and objective items when effect of writing appearance is partialled out of the essay scores

[2] = r between essays and objective items when effect of writing style is partialled out out of the essay scores

[3] = r between essays and objective items when effect of writing is not partialled out of the essay scores

analysis suggesting that the essay scores are not heavily influenced by writing. The lack of consensus between the methods used to ascertain the effect of writing ability on scores yielded by both tests invites further research on this topic.

Students' and Teachers' Opinions

Information gleaned from students and teachers involved in the study suggested that though few people were totally satisfied with the format and expectations of the present ICE, nobody wished to change the format to include only objective items. Though students, in their answers to a forced-choice question, preferred the objective test by a 3 to 1 majority, discussions with many of them revealed that they felt that some essays should be retained on the ICE in addition to the inclusion of multiple-

choice items. This view was shared by their teachers who felt that multiple-choice items were preferable to essays for assessing students' knowledge and skills in many areas of practical and physical geography and to a lesser extent climatology. Essays were seen as preferable for regional geography. Most teachers felt that any attempt to introduce a comprehensive survey of the entire syllabus using multiple-choice items would necessitate a shortening of the syllabus which they even now regard as too broad to be completed in the time allotted.

Implications for Large-Scale Testing

The Intermediate Certificate Exam is representative in form and function of the majority of achievement tests given to secondary school students in countries that have a centralized system of education. There are therefore some lessons that can be learned from this exam.

Scores obtained on public examinations can be instrumental in helping continuing students decide what courses to pursue in upper secondary education and they are often reviewed by potential employers of students opting to discontinue their schooling after the mandatory leaving age. This study has shown that in the case of ICE geography, the format of the exam has a large influence on how and what teachers teach and students learn. Therefore it is imperative that the exam be based as closely as possible on the domain of skills and knowledge defined in the syllabus and that scores yielded by the instrument be as free from error as possible.

Data presented in this study indicate that the essay and objective tests measure different traits. Specificially what these traits are has not been ascertained, though we can speculate that the essay format measures a general intellectual ability and geography knowledge whereas the objective test measures specific facts and skills

relating to geography. Exams such as the Irish ICE, the British General Certificate of Secondary Education (formerly O-Levels), the French Baccalaureate, and the West German Abitur composed primarily of one item type may well fail to measure some of the skills that could be measured using a mixed essay and objective format.

Contrary to the popular belief, objective questions can be developed to measure not only achievement of basic knowledge but higher-order skills as well. There are certain areas of the geography syllabus, which by common consensus, could in fact be better measured using objective items. Specifically, these are the topics subsumed by practical geography, physical geography and areas of climatology. Further research is needed before we can conclude that other areas of the syllabus could be adaquately measured using an objective test format.

Essay exams measuring achievement typically offer examinees choices about what questions they can answer. Standard measurement thinking would suggest that this is psychometrically inadvisable, an assertion that is borne out by the experience of this researcher in trying to calculate a defensible reliablity coefficient for the ICE. In addition, the detrimental effect of such choice on content coverage by students must now be brought into question, particularly when scores on the exam are of such a high-stakes nature. The obvious course of action is to reduce the choice given to examinees regarding which questions they can answer. Ideally, it should be eliminated and all students should be required to answer the same questions.

In education systems where public examinations are the preferred mode of assessing achievement, scores obtained by examinees generally have life-long consequences such as determining access to and courses of study in higher education. In addition, job opportunities are oftentimes a function of one's exam results. In light of these considerations no effort must be spared in ensuring that such exams be as reliable as possible. Test reliability of the essay instrument examined in

this study could be higher, and this could be achieved by reducing the weight given to essays, moving some of the exam to an objective format instead. This study suggests that the essay is a useful vehicle with which to assess some areas of secondary school curricula, researchers concluding that such items offer practice in writing, creativity and formal communication (Milton, 1979). A major problem looms, however, when it is the sole item type used. Retaining the careful scoring procedure as used for the ICE while at the same time embracing some good old-fashioned advice culled from standard test theory, State Education Departments can confidently employ examinations composed of both essay and objective items to assess student achievement of knowledge and skills in many different subjects.

References

Chase, C. J. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement*, 23, 33-41.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory.* New York: Holt, Rinehart and Winston.

Department of Education. (1984). *Rules and programme for secondary schools 1986/87* . (Government Publication No. PL 3976). Dublin: Stationery Office.

Greaney, V. , & Kellaghan, T. (1984). *Equality of opportunity in Irish schools.* Dublin: Educational Company of Ireland.

Hogan, T. P. , (1981). Relationship between free-response and choice-type tests of achievement: A review of the literature. National Institute of Education (ED). Washington, DC. (*ERIC Document Reproduction Service No. Ed 224 811*)

Madaus, G., & MacNamara, J. (1970) *Public examinations: A study of the Irish Leaving Certificate.* Dublin: Educational Research Centre.

Marshall, J. (1967). Composition errors and essay grades re-examined. *American Educational Research Journal*, 4, 375-385.

Marshall, J. , & Powers, J. (1969). Writing, neatness, composition errors, and essay grades. *Journal of Educational Measurement,* 6(2), 97-103.

Milton, O. (1979). Improving achievement via essay exams. *Journal of Veterinary Medical Education*, 6, 108-112.

Murchan, D. (1989). *Comparison of essay and objective test formats for the measurement of achievement in geography in Ireland.* Unpublished master's thesis, Cornell University, Ithaca, NY.

Tollefson, N. , & Tracy, D. B. (1980). Test length and quality in the grading of essay responses. *Education*, 101(1), 63-67.

Author

DAMIAN P. MURCHAN, Graduate Student, 304 Roberts Hall, Cornell University, Ithaca, NY 14853. *Degrees*: B.Ed, St. Patrick's College of Education, Dublin; MS, Cornell University.